

# Identification of a new, abundant superfamily of mammalian LTR-transposons

Arian F.A.Smit\*

Department of Biology, Beckman Research Institute of the City of Hope, Duarte, CA 91010-0269 and Molecular Biology Section, University of Southern California, Los Angeles 90089-1340, USA

Received December 31, 1992; Revised and Accepted March 19, 1993

## ABSTRACT

**A new superfamily of mammalian transposable genetic elements is described with an estimated 40,000 to 100,000 members in both primate and rodent genomes. Sequences known before as MT, ORR-1, MstII, MER15 and MER18 are shown to represent (part of) the long terminal repeats of retrotransposon-like elements related to THE1 in humans. These transposons have structural similarities to retroviruses. However, the putative product of a 1350 base pair open reading frame detected in the consensus internal sequence of THE1 does not resemble retroviral proteins. The elements are named 'Mammalian apparent LTR-retrotransposons' (MaLRs). The internal sequence is usually found to be excised. Their presence in rodents, artiodactyls, lagomorphs, and primates, the divergence of the individual elements from their consensus, and the existence of a probably orthologous element in mouse and man suggest that the first MaLRs were distributed before the radiation of eutherian mammals 80–100 million years ago. MaLRs may prove to be very helpful in determining the evolutionary branching pattern of mammalian orders and suborders.**

## INTRODUCTION

The most numerous transposable genetic elements in mammals are the short and long interspersed nucleotide elements (SINES and LINES) represented in the human genome by Alu with an estimated 500–900,000 copies and L1 with 100,000 copies, respectively (1, 2). New copies of both types of elements find their way into the genome via reverse transcription of an RNA intermediate, a process called retrotransposition. SINES are less than 500 base pairs (bp) long, are transcribed from an internal RNA polymerase III promoter, have an A rich 3' end, and are derived from structural RNA (1, 3). Full length L1 sequences are 6–7 kilobases (kb) long and may contain two open reading frames (ORFs) that code for products related to retroviral proteins such as reverse transcriptase (4, 5). Neither SINES nor LINES have long terminal repeats (LTRs).

The mammalian genome also harbors a variety of relatively low copy number endogenous proutroviruses, which may have entered the germlines of their animal hosts through retroviral

infection of germ cells, and are now stably integrated, vertically transmitted, and more or less incapable of infection (6, 7). Retroviruses may have evolved from an (LTR-)retrotransposon similar to gypsy in *Drosophila* or Ty3 in budding yeast, which acquired an envelope protein gene around the time of the emergence of mammals (8–11). Characteristic of (LTR-)retrotransposons and proutroviruses are two directly repeated sequences of several 100 bp (the LTRs) flanking a central region with more or less preserved ORFs related to the retroviral *gag*, *pol-int*, and sometimes *env* genes (Figure 1). Another hallmark of these transposons is a 4 to 6 bp target site duplication upon integration (of specific length for each type of element). The LTRs are essential and sufficient for normal integration into the host genome; their terminal sequences are recognized by a type-specific integrase, resulting in the exclusive utilization of viral DNA termini for integration (12). Furthermore, the LTRs control all aspects of transcription. LTRs of even closely-related retrovirus families show no overall sequence homology, but all retrotransposon LTRs share short elements functional in integration and transcription: (i) a terminal 5' TG and 3' CA dinucleotide, often extended to a short inverted repeat, (ii) RNA polymerase II promoter elements and transcription start site, and (iii) a polyadenylation signal and site. The transcription start- and polyadenylation sites define the borders between the so-called U3, R and U5 regions in the LTR. Solitary LTRs of endogenous retroviruses in the genome are thought to be excision products of homologous recombination between both LTRs. There are an estimated total of 1100–1600 and 3000–4000 copies of endogenous proviruses and their solitary LTRs in the human and mouse genome, respectively (6,7).

The estimated 40,000 copies of THE1s and their solitary LTRs formed the most widespread interspersed elements known in the primate genome apart from Alu and L1. A considerable number of other repeat families exists in mammals, exemplified by the 21 recently described medium reiterated sequences (MERs) in the human genome (17, 18). The most abundant of these MERs, as determined with a plaque hybridization assay of a genomic human library, is MER18 with 5000–10,000 copies, closely followed by MER10 with 4000–8000 copies (18). The MER10 sequence had already been known to be repetitive (19, 20) and had been named MstII repeat by Mermer *et al.* (21). These authors also had recognized the similarity of these elements to

\* To whom correspondence should be addressed at: Department of Biology, Beckman Research Institute of the City of Hope, Duarte, CA 91010-0269, USA

THE1-LTRs. An alignment of some members of these two (sub)families has been published recently (22). Members of the THE1/MstII family have also been called 'low-repeat sequence' (LRS) (23, 24). It will be shown here that the MER18 sequence, previously described as a human sex-chromosome-specific repeat (25), forms part of the LTR of a retrotransposon related to THE1 and MstII.

The most common interspersed repetitive element described in the mouse genome is L1 followed by the Alu-equivalent B1 SINE, the B2 SINE, and the 'Mouse Transposon' (MT, 26, 27). The latter three have been estimated to occur in similar numbers (1,26). It is shown in this paper that MT is related to the recently partly-described 'Origin-Region Repeat' (ORR-1) in rodents (28) and that both are more distantly related to the primate elements mentioned above. Indeed, several ORR-1 and MT repeats flank sequences that resemble the internal sequence of THE1.

The above mentioned elements, which comprise all of the most common unclassified interspersed repeats in primates and rodents, are identified here as members of a superfamily of Mammalian apparent LTR-retrotransposons (MaLRs). They form a class of mobile genetic elements distinct from SINEs, LINEs, and retroviruses. It is estimated that there are 40,000 to 100,000 copies, including solitary LTRs, in both primate and rodent genomes. I have derived novel consensus sequences for the LTRs of 20 MaLR subfamilies, based on the alignment of over 300 sequences found in GenBank® release 71. These sequences and their putative evolutionary relationship are presented in this paper.

## METHODS

Databank searches were performed on a Sun computer using the IFind (29) program in the IntelliGenetics™ Suite. Multiple alignments were initially made with the Genalign program (30) and significantly adjusted manually. Improved versions of consensus sequences were successively used for new databank searches. Subfamilies were detected when members of a family showed more similarity to each other than to their preliminary consensus sequence or after grouping sequences that share an insertion or deletion. Subfamily status was accepted when a subdivision of a family was accompanied by grouping of consensus sequences with multiple 'diagnostic' deletions, insertions or mutations. Some new subfamilies were detected by searching the databases with sequences that showed an overall (full length) but faint similarity to a previously determined consensus sequence.

For calculation of nucleotide substitution rates, each insertion or gap has been counted as a single substitution. Hypermutable CpG sites were excluded. All sequence divergence or similarity

values mentioned in this article are corrected for superimposed substitutions using the algorithm of Jukes and Cantor (31).

## RESULTS AND DISCUSSION

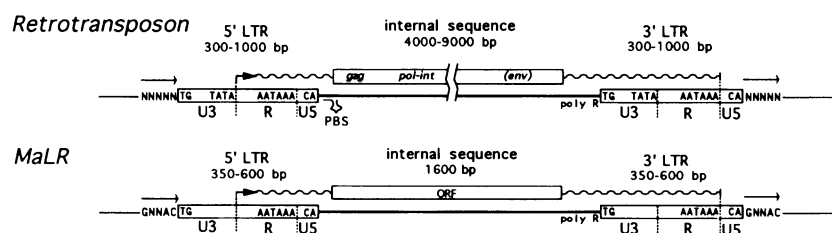
### Identification of a superfamily of LTR-transposons

Initial computer searches were performed to determine the extent of the ORR-1 sequence in the origin of replication region near the Chinese hamster dihydrofolate reductase gene (28). Similarities were detected with MT repeats, and comparison with these elements allowed determination of the exact ends of ORR-1, deleting 30 bp of the 5' end and adding 200 bp to the 3' end of the published consensus (28). Surprisingly, searches with the new full-length ORR-1 consensus showed similarities to several primate sequences, most of which turned out to be THE1-LTRs or MstII repeats. Similarities were subsequently discovered between MstII and both the MER15 and MER18 sequences (18). Through comparison with the MstII consensus, I found that MER15 and MER18 actually represent part of the 5' and 3' arm, respectively, of one element. Consensus sequences of all the elements mentioned could be extended to include 5'-TG and 3'-CA terminal dinucleotides typical for retrotransposon LTRs.

The databanks were also screened with a consensus of the internal sequence of THE1 (adjusted from ref. 32) excluding any LTR sequence. Sequences similar to it were found to be flanked by MstII and MER15/18 elements, and, more surprising, by ORR-1 elements in the Syrian hamster  $\mu$  class glutathione S-transferase gene (HAMMGLUTRA, 33) (Figure 2). For all locations of MaLRs, indicated by their GenBank® locus name in parentheses, refer to Table 1. This is strong evidence for a relationship between the rodent and primate repeat families such as was predicted by the LTR-sequence similarities. Searches with the 500 bp available of the HAMMGLUTRA internal sequence revealed six more internal sequences flanked by an ORR-1 and one by an MT (RATCYPOXG, 34). An element with two ORR-1-LTRs present in the rat cytochrome P450 4A1 gene intron 4 (RATCYP4A1, 35) has a total length of 1912 bp (excluding an integrated B1 repeat) comparable to that of THE1 (2.3 kb).

A third line of evidence for their kinship is that solitary LTRs and complete members of each family are almost always flanked by a 5 bp direct (often imperfect) insertion repeat (see Table 1, column c), as has been observed for THE1 (15). Moreover, the published target site sequence specificity of THE1 (GYNAC) (15) is also obvious for all the other elements (unpublished data).

A picture has emerged of a large superfamily of THE1-like transposons that unites at least six very abundant mammalian repetitive elements: ORR-1 and MT elements in rodents, and THE1, MstII, and MER15 and MER18 in primates. For clarity



**Figure 1.** Comparison of the structure of a typical retrotransposon and MaLR. Noteworthy differences are the short internal sequence and the integration specificity of MaLRs, and the absence of homology to a conventional transcription initiation site, reverse transcriptase, or primer binding site (PBS). See text for details.

of reference, these names will still be used in this article to specify (members of) each family, except that the family of repeats comprising MER15 and MER18 is named MLT1 (Mammalian LTR-Transposon 1). In the future, it may be better to rename the other families MLT2, MLT3, and so on. Alignment of over 300 LTR sequences allowed subdivision of each family, based on the presence or absence of gaps or inserts and multiple diagnostic point mutations (alignment data to obtain subgroup consensus sequences are not shown). 17 of the derived subfamily consensus sequences are presented and compared in Figure 3. The subfamilies are indicated by a small case letter after the family name (e.g. THE1a), with subfamily 'a' being the most recently amplified (see below). Consensus sequences of three ancient MLT1 subfamilies (MLT1e-g), most similar to MLT1d, were too indefinite to be integrated in the Figure 3 alignments.

A total of 311 THE1-related sequences were discovered in the GenBank DNA sequence database (release 71) and are listed in Table 1. Only 30 of these show similarities to an internal sequence, out of which 4 had been isolated by screening with an internal sequence-containing probe. Hence, most MaLRs seem to remain in the genome as solitary LTRs, probably as a result of internal recombination. The LTRs range in length from 327 (ORR1a) to 568 bp (MLT1e). Their terminal 100 nucleotides are relatively well-conserved between families, while the central region is highly divergent in sequence and length. No obvious and conserved potential transcription start site could be located, although a possible TATA-box is indicated in the THE1 and MLT1 sequences in Figure 3. A transcription start site is tentatively positioned 23 bp downstream, supported by sequence information of a processed pseudogene (HUMIGLAB, 36) that apparently had been transcribed from this position (unpublished data). Deka *et al.* (15) suggested a transcription start 40 bp more downstream based on the truncation of a THE1 element at this

point. Notably conserved between all families is the 3' terminal region that contains the polyadenylation signal {AA(T)AAA} and site. This site is usually at a C/TA dinucleotide followed by GT clusters (37), which are both present in each MaLR consensus sequence.

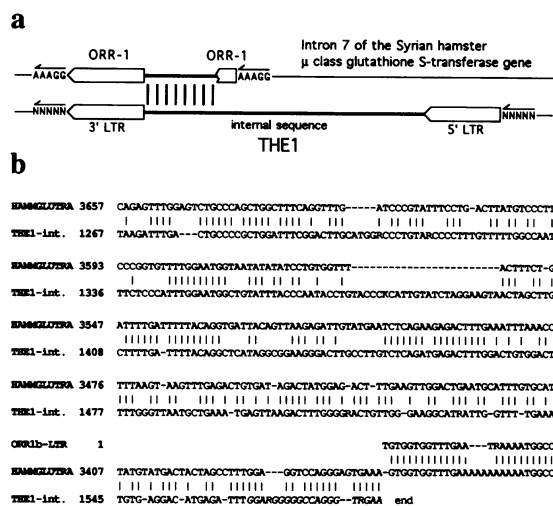
The orientation of the sequences is opposite to the previously published, partial consensus sequences of MstII (21, 22), MER15 and MER18 (18), ORR-1 (28) and MT (26, 27), but conforms to that of the published THE1 sequence (13). It is supported by the presence of a 1353 bp ORF in this orientation in the 1576 bp consensus internal sequence of THE1 elements (unpublished results). Preliminary analysis has not yet revealed significant similarities of the putative product of this ORF to any protein present in the databanks. The present orientation is also supported by 12 cases of transcriptional 3' processing at the proposed site in LTRs of each family (Table 1 and ref. 14). A survey of the orientation of MaLRs within transcription units reveals a very marked (7:1) bias against fixation of positively oriented elements in introns, while no bias in orientation is observed in flanking regions of genes (Table 2). This can be explained by the potential for 3' processing by the LTRs of integrating MaLRs. Integration in the positive orientation inside a gene must have usually led to a premature transcription termination. Selection against alleles with such a mutation is obviously strong.

Reverse transcription of the minus strand of most (LTR-)retrotransposons is primed from a tRNA annealed to a primer-binding site, a short region of complementarity immediately downstream of the 5'-LTR-internal domain junction. There is no complementarity to any tRNA in the consensus internal sequence of the THE1/MstII family, but conventional primer-binding sites are, for example, also absent in the yeast retrotransposon Tf1 (38) and the hepatitis B virus genome (39). The retrotransposon plus-strand is primed at a short polypurine tract just upstream of the internal-domain-3'LTR junction. Consistent with this, 17 of the 20 3' terminal nucleotides in the consensus primate as well as rodent internal sequences are purines (*italicized* in Figure 2).

The structure of the LTRs, the presence of the functional polyadenylation site, the long ORF, the purine-rich stretch, and the 5 bp target site duplication suggest a classification of these elements among (LTR-)retrotransposons and proutetroviruses (Figure 1). However, the term retrotransposon has been reserved for elements with a reverse transcriptase-encoding region, which is seemingly absent in these elements. The name Mammalian *apparent* LTR-Retrotransposon or otherwise Mammalian LTR-Retrosequence (both MaLR) is therefore proposed for this superfamily. Its evidently successful strategy of distribution, apparently without a self-encoded reverse transcriptase, forms an intriguing unknown.

### Evolutionary relationship of the MaLR families

The consensus sequences of MstII and THE1-LTRs show a gradual transformation from THE1a to MLT1a (see Figure 3) that coincides with a gradual increase in average sequence divergence of the copies from their subfamily consensus sequence. A similar correlation can be observed for the MLT1, ORR-1 and MT families. The older (more diverged) subgroups' consensus sequences actually form intermediates between the 'younger' subgroups of different families; for the rodent families MT and ORR-1 highest similarity is seen between MTd and ORR1d, and among the rodent-specific subfamilies the oldest (ORR1d) shows the highest similarity to the MLT1 and



**Figure 2.** Evidence for a relationship between the rodent ORR1 and primate THE1 repeat families. A region in intron 7 of the Syrian hamster  $\mu$  class glutathione S-transferase gene (HAMMGLUTRA, 33) shows homology to the consensus THE1-internal sequence bordered by two ORR-1s. **a**) Schematic comparison of the intron 7 sequence with the primate THE1. An internal deletion apparently has taken place in the hamster element. The remainder is flanked by identical 5 bp repeats, as is observed for all THE1 elements. **b**) Sequence alignment of the intron 7 sequence, in inverse orientation, with the THE1a internal and the ORR1b-LTR consensus sequences. Note that the ORR1b similarity is at exactly the same position as the THE1-3' LTR would be.

Table 1. Location of all MaLR sequences detected in GenBank release 71, ordered by their locus name

locus name	location	a)	b)	c)	d)
<b>PRIMATES</b>					
AGMORS3A	177-542	+	THE1b	TCATG	African green monkey origin-enriched DNA
AGMORS4A	1-242	-	MSTa		African green monkey origin-enriched DNA
CEBGLGBIN	11088-11433	+	MLT1a	GATGR	Chimpanzee A and G gamma globin genes, 3' FR
GAL6 †	1-369 (5' LTR)	+	MSTb		Galago single copy genomic sequence homologous to THE-1.
GAL7 †	15-379 (5' LTR)	+	MSTb		Galago single copy genomic sequence homologous to THE-1.
GCRHBEGEB	14430-14825	+	MLT1a	GATRM	Galago gamma globin gene, 3' FR
GIBHBGGL	9454-9778	-	MLT1a	GATGR	Gibbon A and G gamma globin genes, 3' FR
HUM7SKP17	499-end	+	MLT1a		Human 7SK pseudogene integration site
HUMALMBG1	269-686	-	MLT1c	AACAC	Human $\alpha_1$ -microglobulin-bikunin, 5' FR (-1285 to -870)
HUMACTG3PS	1-75	+	MSTa		Human gamma-actin like pseudogene insertion site.
HUMACTSA	534-901	-	THE1c	CCAGG	Human vascular smooth muscle $\alpha$ -actin, 5' FR (-1213 to -846)
HUMACTSG2	11590-1929	-	MLT1b		Human enteric smooth muscle $\gamma$ -actin, intron 1
HUMADAG	90->230	-	MLT1g		Human adenosine deaminase gene, 5' FR
HUMADAG	1025-  Alu   -1717	+	THE1b	GCCAC	Human adenosine deaminase gene, 5' FR, (-2911 to -2219)
HUMADAG	14448-  Alu   -15469	-	MLT1c	GTYAG	Human adenosine deaminase gene, intron 1
HUMAIGRA	1-end	-	MSTa		Human interspersed repetitive DNA detected with H-ras VTR
HUMAIGRB	1-end	-	MSTa		Human interspersed repetitive DNA detected with HUMSATMA
HUMAIGRC	1-end	-	MSTa		Human interspersed repetitive DNA detected with HUMSATMA
HUMAIGRD	Alu   50-end	-	MSTa		Human interspersed repetitive DNA detected with HUMSATMA
HUMAIGRE	1-end	-	MSTb		Human interspersed repetitive DNA detected with HUMSATMA
HUMALUA	686-1098	+	MLT1b	YATTC	Human genomic clone containing Alu repeat
HUMALURC	Alu   91-255	-	MLT1c		Human clone containing Alu repeat
HUMAMD01	14098-4305	-	MTc		Human S-adenosylmethionine decarboxylase gene, intron 1.
HUMAMINONA	1-163	+	MLT1d		Human aminopeptidase gene, 5' FR (-1071 to -908)
HUMAPB03	1418-1779	-	THE1b	CYRTC	Human apolipoprotein B gene, intron 4.
HUMAPOA1A	<441-899	-	MLT1d		Human apolipoprotein A-I gene, 5' FR (<-2030 to -1586)
HUMBCR22I	1-80   Alu   370-523	-	MLT1b		Human BCR gene, bcr2 region in intron 1
HUMBCR22I	2979-end	-	MLT1		Human BCR gene, bcr2 region in intron 1
HUMBCRI	12-end	-	MLT1a		Human BCR gene, bcr3 region in intron 1
HUMBFXIII	11340-11829	+	MLT1d	TCAGT	Human factor XIII $\beta$ subunit gene, intron 5.
HUMBFXIII	12011-ca.12585	-	MLT1e		Human factor XIII $\beta$ subunit gene, intron 5.
HUMBFXIII	16512-16873	-	THE1b	GCYAC	Human factor XIII $\beta$ subunit gene, intron 5.
HUMBKVH01	1-84	+	MSTb		Human BK virus enhancer like sequence, 5' adjacent region
HUMBLASTIA	261-577   Alu   894-963	+	THE1c	GAC(T)A	Human EBV inducible BLAST-1 gene, 5' FR (-1317 to -615).
HUMC21DIA	1340-1831	-	MLT1c	YAWGC	Human ring chromosome 21 DNA.
HUMCALP09	444-857	-	MLT1b	GCKMG	Human calpastatin gene, intron 10.
HUMCALRT2	1-278	-	MLT1c		Human calretinin gene, intron 1.
HUMCAM3X01	1-86   Alu	-	MSTb		Human calmodulin III gene, 5' FR (<-1370)
HUMCAPG	3050-3364   Alu	-	MLT1a		Human cathepsin G gene, 3' FR
HUMCCG1	15109-end	+	ORR1b		Human X-linked CCG1 cDNA, 3' UTR
HUMCCND3PS	54-407	-	THE1b		Human pseudo-D3-type cyclin gene, 5' flanking site
HUMCETP2	1-348	-	MLT1d		Human cholesteryl transfer protein gene, intron 2.
HUMCFXI1	1-65	+	MLT1e		Human coagulation factor XII gene, 5' FR (<-356 to -292)
HUMCGMP4	2745-ca.3055	-	MLT1a		Human cGMP phosphodiesterase gene $\beta$ -subunit, intron 3
HUMCHR10F	1-end	-	MLT1c		Human chromosome 10 polymorphic microsatellite site.
HUMCLGNA	1974-end	+	MLT1c		Human neutrophil collagenase cDNA, 3' UTR
HUMCOL2C2	<500-end	-	MLT1c		Human type VI alpha-2 collagen gene, 3' UTR
HUMCR2CD21	1-ca.232	+	MLT1f		Human complement receptor 2 gene, 5' FR (to ca.-1100)
HUMCYP450	2710-2978	-	MSTb		Human cytochrome p450 CYP1A1 gene, intron 1
HUMDNAPOLA	1-275	-	MLT1a		Human DNA polymerase $\alpha$ gene, 5' FR (to -1296).
HUMDYSIN7	1810-1972/1007-794	+	THE1b	GATTC	Human muscular dystrophy gene, intron 7
HUMDYSKW	ca.515-868	-	MLT1b		Human muscular dystrophy gene, intron 44.
HUMDYSTRO	12241-12588	-	MLT1a	RTYAC	Human muscular dystrophy gene, intron 44.
HUMDYSTRO	ca.14779-end	+	MSTb		Human muscular dystrophy gene, intron 44.
HUMDYSTROP	22772-  L1   -23666	-	MLT1b	GGRRR	Human muscular dystrophy gene, intron 44
HUMDYSTX60	170-536	-	MLT1a	STCTM	Human muscular dystrophy gene, intron 59
HUMEHS2A	1052-  L1   -2180	-	MLT1e	ATTCT	Human DNA with some similarity to HIV-1 DNA
HUMEXCISR	1514-1501 (int)	+	THE1b		HeLa cell uniform extrachromosomal circular DNA
	1502-end (3' LTR)				
HUMFOLP12	ψ   112-end	-	MLT1b		Human DHFR pseudogene psi-1 integration site.
HUMG6PDGEN	ca.18500-18833	-	MLT1		Human glucose-6-phosphate dehydrogenase, 3' FR
HUMGPP3A01	474-end	-	MLT1a		Human platelet glycoprotein IIIa, intron 1.
HUMGUSBA	735-76   2xAlu   1543-63	+	MSTb		Human $\beta$ -D-glucuronidase gene, 5' FR (ca -3000)
HUMGYPA09	1-310	-	MSTa		Human glycoporphin A gene, 3' UTR.
HUMHAHEP	119-296   Alu   584-850	+	MLT1b	RCYAC	Human Na <sup>+</sup> /H <sup>+</sup> exchanger gene, 5' FR (-1260 to -528)
HUMHAIGGR	7784-8307	-	MLT1e	CWSCA	Human IgG receptor gene, intron 3
HUMHBB	36999-37336	+	MLT1a	GATGR	Human A gamma globin gene, 3' FR
HUMHBB	41777-42114	+	MLT1a	GATGR	Human G gamma globin gene, 3' FR
HUMHBVINT	1259-end	-	MLT1a		Human hepatocarcinoma insertion site of hepatitis B virus
HUMHLAEHCM	1015->1172 (5' LTR)	+	MSTc		Human MHC class I HLA-E/HLA-6.2 gene, 3' FR
	<1218-end (int.)				
HUMHLASBA	8613-9027	-	MSTb	TGGAY	Human MHC class II HLA-SB- $\alpha$ gene, intron 4
HUMHLASBA	10519-10555	+	MSTb		Human MHC class II HLA-SB- $\alpha$ gene, intron 4
HUMHPP16C	367-end	-	THE1c		Human papilloma virus type 16C integration site
HUMHPRTB	7326-  Alu   -8006	+	MLT1a	GTKCT	Human HPRT gene, intron 1
HUMHPRTB	<33866-34315	-	MLT1e		Human HPRT gene, intron 5
HUMHPRTB	54766-55142	+	MLT1b	YTCAC	Human HPRT gene, 3' FR
HUMIFNAR	22644-  Alu   -23293	-	MLT1e	YCYAW	Human interferon alpha/beta receptor gene, intron 6
HUMIGCC5	1-57 (int) 58-552 (LTR)	-	MSTb		Human inactive IgH C $\mu$ 2 gene, 5' FR
HUMIGCD3	1-57 (int) 58-504 (LTR)	-	MSTb		Human IgH C $\mu$ 1 gene, 5' FR
HUMIGCD3	505-667	+	MLT1b		Human IgH C $\mu$ 1 gene, 5' FR
HUMIGHEC	1-end (5' LTR),	-	MSTb		Human IgH C $\gamma$ 3-C $\gamma$ 1 intergenic region.
HUMIGHHSG3	1-665   (internal)				
HUMIGHHSG4	1-665   (internal)	-	MSTb		Human IgH C $\gamma$ 4- C $\mu$ intergenic region.
HUMIGHVHA	1470-end	+	MLT1b		Human IgH variable region, hv4005 gene, 3' FR
HUMIGHVV2	979-end	+	MLT1b		Human IgH variable region, V71-2 gene, 3' FR
HUMIGKV18B	<388-517	-	MSTc		Human IgH 5VDJ-CG region recombination site
HUMIGLAB	24-88	+	THE1c		Human IgL J-C region processed pseudogene, 5' UTR.
HUMIGLAMB	32835-33393	-	MLT1d	GGRGC	Human Ig lambda constant region 7, 3' FR

HUMIGMUD	3400->3635	-	MLT1a	Human IgH C $\mu$ -C $\delta$ intron/intergenic region.
HUMIGMUD	4638-7324	-	MLT1c	Human IgH C $\mu$ -C $\delta$ intron/intergenic region.
HUMINSRC/D	1180-1 2xAlu   -2395	-	MSTa	Human insulin receptor, intron 14
HUMINT2	11442-end	+	MLT1c	Human int-2 proto-oncogene, 3' FR.
HUMLAC102	<53-end	+	MLT1b	Human lipoprotein associated coagulation inhibitor, exon 2
HUMLDHB/C	1507-1574   $\Psi$   1-end	-	MSTa	Human lactate dehydrogenase B pseudogene integration site
HUMLTR	224-2531	-	THE1c	Human muscular dystrophin gene, intron 43.
HUMMER15A	1-154	-	MLT1b	Human repetitive DNA fragment sequenced from Alu-primers
HUMMER15B	1-141	-	MLT1c	Human repetitive DNA fragment sequenced from Alu-primers
HUMMER18A	78-end	-	MLT1b	Human repetitive DNA fragment sequenced from Alu-primers
HUMMETIPG	195-299   $\Psi$	+	MLT1b	Human methalothionein I pseudogene integration site.
HUMNK25	437-866	+	MSTc	Human neurokinin A receptor gene, exon 5 and 3' FR
HUMNKG2D	<1489-end (int.)	+	MLT1	Natural killer cell membrane protein NKG2D cDNA, 3' UTR
HUMP450SCC	1-214	+	MLT1f	Human CYP11A1 gene, 5' FR (to -2117)
HUMPADP	1-ca 250	+	MLT1b	Human amyloid A4 precursor gene, 5' FR (-3700 to -3450)
HUMPCI	1-281	-	MLT1d	Human protein C inhibitor gene, 5' FR (-2200 to -1919)
HUMPCI	6854-7246	-	MLT1b	Human protein C inhibitor gene, intron 1
HUMPCI	9254-9603	-	MLT1a	Human protein C inhibitor gene, intron 2
HUMPECORIA	1-ca.220	-	MLT1e	Human polymorphic site near <i>THRB</i> gene.
HUMPPB1A2	1->169	-	MSTb	Human protein phosphotyrosyl phosphatase 1B gene, intron x.
HUMPROSA	1-202	-	THE1b	Human vitamin K dependent protein S, 5' UTR
HUMPS12	1991- >2191	+	MSTa	Human clone ps12 DNA (from artificial chromosome?)
HUMPS94A	1-87	+	MLT1b	Human prostatic secretory protein PSP94; 5' FR (to -740).
HUMRFLP3	1-111	+	MLT1b	RFLP site in human genome linked to Huntington's disease.
HUMRSO4C	26-322	-	THE1a	Original o-repeat, genomic clone found to be repetitive
HUMRSO5C	26-365	-	THE1a	Human genomic clone isolated with HUMRSO4C probe
HUMRSOLTR	21-2286	+	THE1a	Human genomic THE1 isolated with o-repeat clone
HUMSAA1A	1954-2374	+	MSTa	Human serum amyloid A, intron 1
HUMSATOD	411-786	+	THE1c	Target site of human chr. 6 duplication unit on chr.16
HUMSERG	4104-4482	-	MLT1b	Human serglycin gene, intron 1
HUMSEXREP3	351-579/580-816/817-end	-	MLT1b	Tandem repeats on human Y pseudoautosomal region
HUMSEXRPA	351-580/581-818/819-end	-	MLT1b	Tandem repeats on human Y-q
HUMSIGM53	1646-2001 (int.)	+	MLT1c	Human IgH constant region C $\gamma$ 3 gene, 5' FR
	2002-2441 (3'LTR)			
HUMSIGM4	2137-2290 (int.)	+	MLT1c	Human IgH constant region C $\gamma$ 4 gene, 5' FR
	2291-2732 (3'LTR)			
HUMSMAAA	1-251	-	MSTb	Human aortic-type smooth muscle alpha-actin, intron 8.
HUMTHEP2	1-136	-	MLT1d	Human genomic integration site of a THE1.
HUMTHEP2	158-493   Alu   783-824	+	THE1b	Human genomic clone isolated with probe for o-repeat
HUMTNF2	5249-5602	-	MLT1d	Human tumor necrosis factor receptor, 3' FR
HUMTP001	697->1076	-	MLT1d	Human thyroid peroxidase gene, 5' FR (-1903 to ca. -1500)
HUMTP004	1-448 (int) 449-866 (LTR)	+	MSTb	Human thyroid peroxidase gene, intron 3
HUMTP006	1790-2231 (3'LTR)	+	MLT1d	Human thyroid peroxidase gene, intron 6
	2232- >2610 (int.)			
HUMTRANSC	498-2699	-	THE1a	Human calmodulin family member gene, 3'-UTR
HUMUBIBP	1-106   $\Psi$   595-779	-	MLT1e	Human ubiquitin Ub B pseudogene integration site.
HUMUDPCNA	L1   315-519	-	MSTc	Human UDP--acetylglucosaminyltransferase I gene, intron 1
HUMUG2PC	261-end	-	MSTb	Human U2 small nuclear RNA pseudogene insertion site.
HUMUG4PB	1-162   $\Psi$	-	THE1b	U4 pseudogene integration site, 140 bp 5' of coding region
HUMVCAMA	1-52	-	THE1a	Human vascular cell adhesion molecule-1 gene, 5'FR(<-2948)
HUMVWFAB	1-113   $\Psi$	+	MLT1d	Human von Willebrand factor pseudogene recombination site
HUMXTO....				Sequenced tags of human brain cDNAs. There are similarities to 11 THE1-, 9 MstII-, and 27 MLT1-LTRs and 3 internal sequences.
M24686	1-200	-	MLT1b	Human angiotensinogen (serpin) gene, intron 1
MACGLINE	9380-9722	+	MLT1a	Rhesus monkey A and G gamma globin genes, 3' FR
MACPSP94	1-122	+	MLT1b	Rhesus monkey prostatic secretory p94, 5'FR (to -740)
MACRHPV1	371-end	+	MSTb	Integration site for Rhesus monkey papilloma virus
MNKGLINE	5370-5721	+	MLT1a	Spider monkey A and G gamma globin genes, 3' FR
TARHBEGPS	10694-11048	+	MLT1a	Tarsier gamma globin gene, 3' FR

**RODENTS**

CRUDHFRORI	3112-3364	-	ORR1b	GTAGY	Chinese hamster DHFR gene origin region of replication
CRUDHFRORI	5357-5396	-	MLT1a		Chinese hamster DHFR gene origin region of replication
CRURSA49C	1-30   Alu   1216-end	-	MTd		Chinese hamster Alu type 2 integration site.
HAMADEP	<980->1200 (int.)	+	ORR1b		Syrian hamster androgen-dependent expressed protein cDNA, 3' UTR.
	1335-end (3'LTR)				
HAMCADA	Mys   586-634	+	MTC		DNA recombined with <i>CAD</i> gene 3'FR upon CAD amplification
HAMMGLUTRA	3003-3369 (3'LTR)	-	ORR1b	CCTTT	Syrian hamster $\mu$ class glutathione S-transferase gene, intron 7.
	3687-3754 (5'LTR)				
HAMV11	24-53 (LTR) 54-end (int)	+	ORR1b		Adenovirus integration site in hamster kidney cell line.
M22992	2544-2850 (5'LTR)	+	ORR1a		Mouse $\alpha$ -subunit gene of thyrotropin, intron 3
	2851-2992   B1 (int)				
M25490	1-110	-	MTC		Rat bone gla protein (BGP) gene, 5. FR (to -500)
MUSAB321	724-1073	+	MTb	AGTAC	Recombination hotspot in mouse MHC <i>Pb/Ob</i> intergenic region
MUSACHRA	524->918	-	MTd		Mouse acetylcholine receptor $\gamma$ subunit, 5'FR(-2519 to-2127)
MUSACHRBB	4449-4775   B2	-	MTd		Mouse acetylcholine receptor $\beta$ subunit gene, intron 7
MUSAGRIN6	526-700	-	ORR1b		Mouse agrin gene, 3' UTR
MUSANTP91A	5638-5954	-	ORR1c	GRCCC	Mouse transplantation antigen P91A, intron 5
MUSANTRVL1	1-16   IAP	+	MTd		Integration site of IAP DNA 9 kb upstream of the angiotensinogen gene in Swiss mice
MUSANTRVL2	IAP   403-454				
MUSAP1	3520-end	+	ORR1a		Mouse alkaline phosphatase pseudogene, 3' FR
MUSB7BLAA	1239->1442	+	MTd		Mouse B lymphocyte activation antigen B7 mRNA, 3'UTR
MUSBFCGR1	9749-10085	-	ORR1b	CCTTT	Mouse beta-Fc-gamma-RII gene intron 4
MUSBGCXD	34929-   L1   -36733	-	MTC	AGYTC	Mouse $\beta$ -globin bh3-b1 intergenic region.
MUSBPP9	2281-2601	-	MTE	CTKAS	Mouse single stranded-DNA binding protein cDNA, 3' UTR
MUSCR2A	2219-2314 (5'LTR)	+	ORR1b		Mouse complement receptor cDNA, 3' UTR
	2480-end (3'LTR)				
MUSCYTCC	1486-1856	+	MTb	ACART	Mouse <i>Cyp2d-11</i> gene, intron 1
MUSDALH2K	795-967   B2	-	ORR1a		Mouse B1 and MT repeat DNA region
MUSDALH2K	1106->1265	-	MTC		Mouse B1 and MT repeat DNA region
MUSEB2INTR	2926-3266	-	MTC		Mouse MHC class II E $\beta$ gene, intron 2.
MUSECDD	1-53   L1   372-end	-	ORR1b		Mouse thymus extraction circular DNA
MUSEF2PS	64-459	+	Mta	GGCTG	Mouse elongation factor 2 processed pseudogene, 5' FR
MUSERPRA	1-58	-	ORR1b		Mouse erythropoietin receptor gene, 5' FR (to -1640)
MUSF9GP	237-end (-Py)	+	Mta		Mouse DNA joined to polyoma virus, integrated in rat genome
MUSFBP2A	1-101	+	ORR1b		Mouse folate-binding protein 2 cDNA, 5' UTR

MUSIGCD17	2962-3347	+	ORR1c	RGGM	Mouse IgH constant region C $\delta$ gene, 3' UTR
MUSIGCR	<1925-2174	-	MLT1		Mouse IgH constant region C $\gamma$ 2a gene, 3' FR
MUSIGMUD3	3554-39071	-	MLT1c		Mouse IgH constant region, C $\mu$ -C $\delta$ intergenic region.
MUSIGMUD3	13908-4063	+	MLT1c		Mouse IgH constant region, C $\mu$ -C $\delta$ intergenic region.
MUSIL2R	B1  668-802	-	MTc		Mouse interleukine receptor, 5' FR (-800 to -650)
MUSINSO1	250-657	-	MTb	RGATC	Polyoma virus integration site in mouse genome
MUSLDF	3922-4033	-	MTa		Mouse formin mRNA, alternatively spliced coding exon.
MUSLDHAG	9923-ca.10266	-	ORR1d		Mouse lactate dehydrogenase-A, intron 5.
MUSLDHAG	<12704-end	-	ORR1b		Mouse lactate dehydrogenase-A, 3' FR
MUSLDHAPS	2460-2703	-	ORR1c		Mouse lactate dehydrogenase-A processed pseudogene, 3' FR
MUSLFP	351-490	-	MTc		Mouse lactoferrin gene, 5' FR (-2310 to -2170)
MUSMCKA	3219-3551	-	ORR1d	GGSTC	Mouse muscle creatine kinase gene, intron 1
MUSMHC4H2S	7869-8265	-	MTa	CACAC	Mouse MHC complement component C4 gene, intron 15.
MUSMHCAB1	(TG)n-90-498	+	MTd		Mouse MHC class II Ob gene cDNA, 5' FR (-1180 to -760)
MUSMHC2A	<4820-5425 (int.)	+	ORR1c		Mouse MHC class I thymus leukemia antigen (Tla)-T2a-a chain (H-2a) gene, 3' FR
MUSMHC2A	5426-end (3'LTR)	+	ORR1c		Mouse maternally transmitted MHC class I gene, intron 3
MUSMHH2CAS	2028-2425	-	MTc	GT(T)AG	Mouse MHC class I gene H-2Kb, 5' FR (to -1697)
MUSMHHBA	1-323	+	ORR1b		Mouse MHC class I Tla gene 17.3.A, 3' FR
MUSMHTLAC	5157-end (int.)	+	int.		Mouse multifinger gene mKr2 cDNA, 3' UTR
MUSMKR2	2257-end	+	MTc		Mouse HyperMutable minisatellite locus
MUSMS6HM	37-511	-	MTa	GTYAG	Mouse genomic clone selected with MT repeat in MUSINS
MUSMTREPA	1-end	-	MTa		Mouse genomic clone selected with MT repeat in MUSINS
MUSMTREPB	1-end	-	MTb		Mouse genomic clone selected with MT repeat in MUSINS
MUSMTREPC	1-end	+	MTb		Mouse DNA fragment contacting the nuclear lamina
MUSNLAM23	1-end	-	MTc		Mouse mastocytoma proteoglycan core protein gene, intron 2
MUSPCPA2	1084-1140  B2	-	ORR1a		Mouse hexamer repeat sequence genomic clone
MUSPERIOE	475-end	-	ORR1a		Mouse spleen mRNA containing period repeat, 5' UTR.
MUSPERSP5	1-157	+	ORR1b		SV40 transformation induced polIII transcribed B2 RNA
MUSPOLRSB2	73-413-polyA	+	MTa		Mouse proline rich protein (M14) gene, 3' FR
MUSPRPC2	6992-end	-	MTb		Mouse MHC class I (Qa) Q2-k gene, 5' FR (-552 to -236)
MUSQ2K1	1-317	+	MTd		Mouse MHC class I (Qa) Q1-k gene, 5' FR (-540 to -158)
MUSQAQ1K	1-282	+	MTd		Mouse renin gene, 3' FR
MUSREN2IA	1-97  IAP  3101-3289	+	MTc		Mouse genomic clone detected with MUSREPMT2a
MUSREPMT1	173-end	-	MTa		Mouse genomic clone detected with MUSREPMT2a
MUSREPMT2	263-end	-	MTa		Mouse cerebellum mRNA repetitive portion
MUSREPMT3	12-end	-	MTa		Mouse genomic clone detected with MUSREPMT2a
MUSREPMT4	188-end	+	MTa		Mouse ribosomal protein L32' processed pseudogene, FR
MUSRPL32A	2719-3048	+	ORR1a	GTGAC	Mouse sequence flanking di-2 repetitive element.
MUSRSDI2A	548-end	+	MLT1a		Mouse serum amyloid A-1 gene, 3' FR
MUSSAA1B	3675-3740	-	MTc		Mouse serum amyloid A-2 gene, 3' FR
MUSSAA2B	3624-3919	-	MTc		Mouse U3 processed pseudogene, 5' FR
MUSSNU3P	74-173<	+	ORR1a		Mouse t-complex gene Tcp-1x cDNA, 3' UTR
MUSTCP1X	1600-1680-polyA	+	MTc		Mouse thrombomodulin gene, 5' FR (-2600 to -2100)
MUSTHROMBO	ca.171-569	+	MTc		Mouse expressed sequence tag.
MUSTSG64X	1-end	-	MLT1d		Mouse T cell receptor $\alpha$ chain variable region, Va 8 3' FR
MUSTSPCO2	1675-2029	-	ORR1d	GRRAT	Mouse U7 processed pseudogene, 5' FR
MUSU7P523	1-124	-	MTa		Mouse U1 processed pseudogene integration site.
MUSUG1PB	1-110  $\Psi$	+	ORR1a		Mouse ypt1 gene for ras-related protein, intron 2
MUSYPT13	1-172	-	ORR1a		Rat genomic DNA containing multiple repetitive elements
RAT55REP	B2  120-348	+	MLT1		Rat acyl-peptide hydrolase gene, intron 14
RATAPI	8684-8985	-	ORR1b	KCAAY	Rat NADH-cytochrome b5 reductase gene, 5' FR(to -1100)
RATB5RG	1-278	+	ORR1d		Rat catalase gene, 5' FR (-4040 to -3630)
RATCATO1	634-ca.1045	+	ORR1d		Rat CEA related protein 1 gene, intron 4
RATCGM1AC6	1262-501	-	MTc		Rat CEA related protein 4 gene, intron 4
RATCGM4AA	4096-4333	-	MTc		Rat gamma A-B crystallin genes, intergenic region
RATCRYG	14783-15373	+	ORR1a	MSACC	Rat gamma A-B crystallin genes, intergenic region
RATCRYG	ca.14910-15195	-	ORR1a		Rat gamma B-C crystallin genes, intergenic region
RATCRYG	22216-22567	-	ORR1b	CCAAY	Rat gamma C-D crystallin genes, intergenic region
RATCRYG	ca.29694-29990	-	ORR1d		Rat gamma D-E crystallin genes, intergenic region
RATCRYG	41224-41627	-	ORR1d	TARRG	Rat CYP2B1 gene, 5' FR (-1242 to -743)
RATCY45E1	264-663	-	MTc	RYTTY	Rat CYP2B2 gene, 5' FR (to -743)
RATCYPPB4	1-77	-	MTc		Rat CYP17 gene, 5' FR (to -1685)
RATCYP17G	<1487-1861	+	MTc		Rat clofibrate-inducible CYP4A1 gene, intron 4
RATCYP4A1	4275-6343	-	ORR1b	RTYAT	Rat clofibrate-inducible CYP4A1 gene, 3' UTR
RATCYP4A1	<14580-15019	-	MLT1d		Rat genomic clone, ca. 4 kb upstream of NADPH-cytochrome P-450 oxidoreductase gene
RATCYPXOG	1-238 (5'LTR) 239-465  DI 577-647 B1 (int)	+	MTc		Rat vitamin D binding protein (Gc-globulin) gene, intron 2
RATDBP02	882-ca.1210	-	ORR1b		Rat insulin-like growth factor IA gene, 3' FR
RATGFIL4	1136-end	-	MLT1		Rat CYP2D3 gene, 3' FR
RATIID3G	8747-9107	+	MTd	CCAGT	Rat plasma kallikrein gene, intron 7
RATKALP09	1-153	-	MTb		Rat plasma kallikrein gene, intron 8
RATKALP09	377-569(TR)n	-	MTd		Rat plasma kallikrein gene, intron 8
RATKALP10	1-60	-	MTd		Rat plasma kallikrein gene, intron 10
RATKALP11	1-118	+	ORR1a		Rat MT "gene", unpublished
RATMT	all 64 bp	-	MTa/b		Rat methalothionein-1 processed pseudogene B, 5' FR
RATMT1PB	377-701	-	ORR1a	RGRGA	Rat s-myc gene, 3' FR
RATMYCS	<7800-end	-	int.		Rat nicotinic acetylcholine receptor-related mRNA, 3' UTR
RATNACHRR	2006-2363	-	ORR1b	TTAAG	Rat ornithine transferase pseudogene integration site
RATOAT3	190-299	+	ORR1d		Rat ornithine transferase pseudogene integration site
RATOAT3	300-582  $\Psi$	+	ORR1d		Rat osteocalcin gene, 5' FR (-762 to -503)
RATOST	333-592	-	MTc		Rat prostatein C3 subunit gene, 5' FR (to -2624)
RATPRSTTNC	1-130	-	ORR1d		Rat renin gene, 3' FR
RATRENA	13352-13717	+	MTc	WMSCT	Rat salivary proline-rich protein (RP15) gene, 3' UTR
RATRPL5	5352-end	-	MTb		Rat DNA associated with synaptonemal complexes.
RATRAT51	1-end	-	MLT1		Rat S-adenosyl decarboxylase gene, 5' FR (-1438 to -921)
RATSADMEDC	2-374  B2  504-518	-	MTd		Rat synapsin I gene, 5' FR (to -1096)
RATSYNIA	1-377	-	ORR1c		Rat TRPM2 gene 5' FR (-4423 to -4088)
RATTRPM2B	1330-1655	+	ORR1a	GWATA	
<b>OTHER MAMMALS</b>					
BOVHIOMT	1441-1770	+	MLT1c		Bovine hydroxyindole O-methyltransferase mRNA, 3'UTR.
BOVOXNC5	AS  600-1127	+	MLT1e		Bovine oxytocin gene, 5' FR (to -2091)
BOVSHIRP	1-143  AS  356-377	+	MLT1d		Bovine Alu-like repeat insertion site.
RABCYTR4	1804->1990	-	MLT1		Rabbit CYP4A7 gene cDNA, 3'UTR
RABLS1B	3244-3311  C  3657<	-	MST/MLT1a		Rabbit 1 S-1 gene for arylamine N-acetyltransferase, 3' FR
SHPBLGA1	1023-1369	-	MLT1	GRGTC	Sheep beta lactoglobulin, intron 1.

THE/MstII families (highest to MLT1a). These observations and the distribution of the (sub)families over mammalian species suggested an evolutionary relationship of MaLRs as depicted in Figure 4.

The significant difference between subfamilies in average sequence divergence of copies to their consensus (Figure 4) is consistent with a punctuated nature of subfamily formation. Similar observations have been made for Alu and L1 (reviewed in 40). The consensus sequence of each subgroup may represent the approximate sequence of one or a few transpositionally competent 'source elements' or 'master genes' at the various periods during evolution when they gave rise to a much larger number of defective elements than in intermittent periods. There is no indication of a contemporary distribution of elements in human, mouse, or rat, although the existence of small groups of recently distributed MaLRs, with too few representatives in the databanks to be recognized, cannot be ruled out. It is interesting to note that the length of the LTRs generally seems to have declined in evolution; the youngest member of each family always has the shortest consensus sequence (see Figure 3).

ORR-1 and MT MaLRs form two families confined to rodents. They are more similar to each other than to the other families and may share a common ancestor in an early rodent. Their occurrence in presumably human sequences can actually be an omen for a cloning artifact. Indeed, the 3' end of a human CCG1 cDNA (HUMCCG1), which contains an ORR1-LTR, was found to be of hamster origin (41). This may also be the case for the sequence including the MT in intron 1 of the human S-adenosylmethionine decarboxylase gene (HUMAMD01, 42), further evidenced by a drop from 12% to 1% in CpG content of the DNA before and after the MT homology. MTa, the most recently amplified MaLR subfamily, has, so far, only been found in mouse sequence entries. This is consistent with its average sequence divergence from the consensus of 6.5%, which is less than half the synonymous divergence between rat and mouse (18–23%) (43,44), indicating that it amplified after the mouse-rat split.

MstII and THE1-MaLRs form a primate branch of the superfamily. The only sequences hybridizing to a human THE1a clone in genomic DNA of the prosimian galago, GAL6 and GAL7 (45), are members of the MSTb subgroup. Comparison of this subgroup's sequence divergence in the human genome (21%) to the estimated divergence of noncoding human DNA since the diversion from prosimians 50–60 million years (Myr) ago (13–19%) (46, 47) supports an amplification prior to this event. Accordingly, the MSTa and THE1 subfamilies have substitution levels supporting a later distribution in simians only.

In contrast, members of the MLT1 family, predominantly found in primate databank entries, are also present in rodent, rabbit, and artiodactyl (cow and sheep) genomes (see Table 1). This family is presumably the oldest group in the MaLR superfamily. The divergence percentage of most MLT1 subfamilies agree with a distribution before primate evolution.

Indeed, Kaplan *et al.* (18) found that their MER18 probe (= MLT1b), but not their MER10 probe (= MSTb, HUMHLASBA) hybridized to bovine chromosomal DNA. However, hybridization to mouse or hamster DNA was not observed. The apparently much higher neutral nucleotide substitution rate in rodents than in higher primates and other mammals (48) may obscure detection of 80–100 Myr old MLT1 elements in rodent genomes both by hybridization or databank searches. This could be an explanation for the relatively low number of MLT1-MaLRs found in the rodent databases and the failure of the MER18 probe to hybridize with rodent DNA, although it is also possible that the major amplification of MLT1 elements occurred after the rodent-primate split.

An MLT1a element is present in the gamma globin region of all studied simians and prosimians (HUMHBB, CEBGLOBIN, GIBHBGGL, MACGLINE, MNKGLINE, GCRGEBEB, TARBGPS) (47) implying that this transposon has integrated over 55 Myr ago in the DNA of a common ancestor of at least all primates. In fact, an orthologous MLT1c-MaLR seems present in the immunoglobulin heavy chain C<sub>μ</sub>-C<sub>δ</sub> intergenic region of both human and mouse (HUMIGMUD, HUMIGCMUDE, MUSIGMUD3) (49–51) (Figure 5a). It is present in the human genome with a full-length internal sequence and two LTRs (the 366 bp repeats in ref. 50). Akahori *et al.* (52) noted that one of two 63 bp repeats (the 'sigma-gamma core sequences'), which are part of the R-U5 region of the MaLR's LTRs, is conserved in the mouse C<sub>μ</sub>-C<sub>δ</sub> intron, leading them to suggest a function for this sequence in immunoglobulin expression or construction. Actually, a 150 bp region in mouse that is inverse duplicated (comprising the 'unique sequence inverted repeats' in ref. 51) is 69% similar to both the MLT1c consensus and the 5' LTR of the human MLT1c (Figure 5b). Several lines of evidence (see legend to Figure 5) suggest that this MaLR has integrated before the diversion of rodents and primates.

The above observations imply that the MaLR class of transposons has originated before the radiation of eutherian mammals 80–100 Myr ago. A much more recent origin had previously been suggested based on  $\phi$ -tests indicating that THE1 like repeats are present as single or oligo loci copies in prosimians and in high copy number in higher primates (53) and on sequence data from the prosimian galago, suggesting that the internal sequence had become flanked by LTRs during simian evolution (45).

It has been suggested that the study of the taxonomic distribution of Alu elements (from 55 Myr ago on) can be used to solve the branching order in the higher primate evolution (3, 54) and the distribution of a rodent L1-subfamily (Lx) has been used to delineate the murine subfamily relationships (55). Since many (abundant) MaLR subfamilies seem to have amplified during the radiation of the eutherian (sub)orders, the detection of the presence of orthologous elements or the general distribution patterns of these elements may be used to untangle this higher order branching pattern. For example, the ORR1b-MaLR in the

Position numbers refer to those in the database entry. | denotes an abrupt end to the homology with the consensus. If this is caused by recombination with or integration of a known element, this is indicated ( $\psi$  = pseudogene, Mys = Mys endogenous proviral LTR, ID = rat identifier element, IAP = rodent intracisternal A particle DNA, AS = artiodactyl SINE, C = rabbit C-repeat). < and > indicate possible extension of the element. (int.) = internal sequence <sup>a</sup>) Orientation of the element in the sequence entry. <sup>b</sup>) Type of LTR as presented in Figure 3. Elements with internal sequences are underlined. <sup>c</sup>) Target site duplication sequence in the orientation of the element. The same symbols for degenerate bases are used as in figure 3. <sup>d</sup>) Description of site. UTR = untranslated terminal region of mRNA. FR = flanking region. † No databank entries exist for GAL6 and GAL7 (45). \* The left and right arm of the THE1b-LTR in HUMDYSIN7 (60) are separated and face opposite directions.

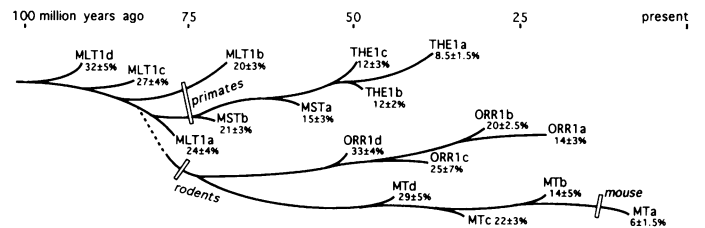


**Figure 3.** Alignment of MaLR-LTR consensus sequences. Each sequence shown is a consensus sequence and defines a subfamily. It is derived by alignment of at least 6 members found in the databases. Grouped consensus sequences represent families. Families could only be aligned in the regions denoted with a gray bar between the grouped consensus sequences. Conserved sites are shown at the top line, with capitals indicating (virtually) invariable sites. The MSTc consensus is partial, since homology extended only between two members beyond the sequence presented. The consensus sequences of three more, highly diverged MLT1 subfamilies are still too indefinite to be integrated in this figure. The underlined region in the U3 part of the ORR1a and ORR1b consensus sequences is often found to be tandemly duplicated. The conserved (and functional) polyadenylation signal and site (the R/U5 boundary) are indicated. Similarly, a tentative TATA-box and transcription initiation site (the U3/R boundary) are marked. The length of the consensus sequences is given at the end of each sequence. R = A/G, Y = T/C, W = A/T, M = A/C, K = G/T, S = G/C, N = A/G/C/T. Underlined numbers in the consensus sequences represent inserts of that length.

Syrian hamster  $\mu$ -class glutathione S-transferase gene (HAMMGLUTRA) that is absent in the same gene in rat. This is due to a MaLR insertion in the hamster lineage rather than to a deletion in the murid lineage, since the apparent deletion in rat DNA (33) comprises exactly the above MaLR sequence plus one of the 5 bp insertion repeats. Since members of the ORR1b subfamily are present in both murids and hamsters, they must have been distributed around the time of the hamster-murid split. Their average sequence divergence is consistent with this. Most rodents more closely related to hamsters than murids could therefore be expected to be 'labeled' with this MaLR insert.

**Estimate of the number of MaLRs in the genome**

Over time, most MaLRs have diverged considerably from their consensus sequence. This, and the existence of multiple subfamilies, complicates estimates of their frequency in the genome by hybridization experiments. For instance, Kaplan *et al.* (18) estimated the number of MER15 and MER18 elements in humans to be 700 to 1,500 and 5000 to 10,000, respectively, although they represent the 5' and 3' arm of the same MaLR-LTR subgroup (MLT1b). The 3' arm is better conserved between MLT1 subgroups, possibly accounting for this discrepancy. Related difficulties are also evident in the original estimates of



**Figure 4.** Schematic representation of the putative relationship of the MaLR families and subfamilies, in part based on their distribution among mammalian species and the sequence alignments in Figure 3. The tip of each branch corresponds to the approximate period of amplification for each subfamily as calculated from the average (corrected) sequence divergence of the copies from their consensus sequence. These divergence values, presented with standard deviation underneath the subfamily names, are for copies found in human DNA, or, for ORR1 and MT, in murine DNA. The time scale functions only as a general guideline, since the correlation of sequence divergence and age depends on disputed assumptions regarding neutral nucleotide substitution rates (52, 58, 59). Values used are  $6.5 \cdot 10^{-9}$  substitutions/site/Myr for rodents (48), and the over evolution gradually diminishing rates for the human branch as calculated by Bailey *et al.* (46).

the THE1-LTR reiteration frequency (16). Based on S1 nuclease protection of their THE1a-LTR (o-repeat) clone by genomic DNA fragments, a frequency of 2,000 and 37,000 elements per human haploid genome was estimated when using stringent or less



**Table 2.** Orientation of MaLR sequences in comparison with the transcriptional unit in or near which they are located

orientation	5' flanking	introns	3' UTR and 3' flanking	intergenic	total
similar	21	8	18	5	52
inverse	18	56	20	7	101

Orthologous elements and elements multiplied through gene duplications have been counted only once. The bias within introns against fixation of elements in the same orientation as the gene is probably due to the presence of the potent polyadenylation site in MaLR-LTRs.

**Table 3.** The number of copies of each family found in GenBank release 71, and estimates of the reiteration frequencies of each family in the human and mouse genomes.

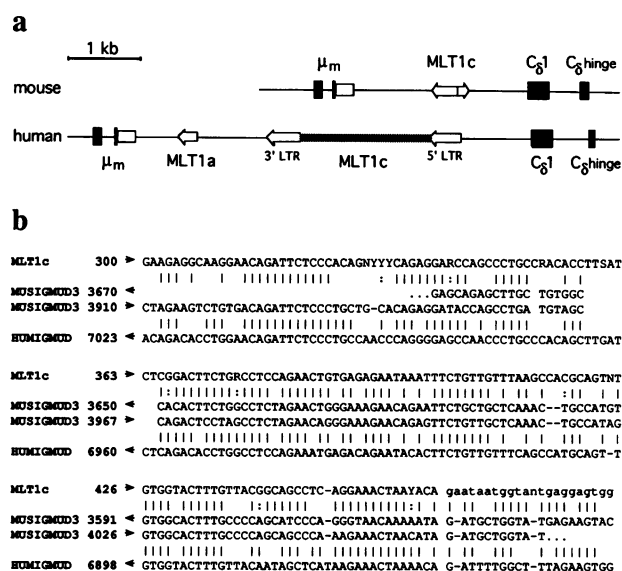
	human databases	genome	mouse databases	genome	rat databases	cow/sheep databases	rabbit databases
THE1	31 (26)	9–16,000					
MstII	40 (36)	12–21,000					
MLT1	101 (101)	34–60,000	4 (4)	22–6,000*	4 (4)	4 (4)	2 (2)
ORR1	1*		25 (25)	10–38,000	18 (18)		
MT	1*		39 (33)	13–50,000	16 (16)		
total	172 (163)	55–97,000	68 (62)	25–94,000	39 (39)		

The numbers between parentheses indicate the number of elements sequenced by chance, i.e. not by searching with a MaLR-probe. These numbers have been used to estimate the relative frequency of each family in the genomes. For the estimations of the absolute numbers I have used a conservative estimate of 500,000 Alu in the human and 80,000 B1 and B2 elements in the mouse genome (1). \* Probably of artificial origin. † Possibly an underestimate since most copies may have diverged too much to be detected.

stringent digestion conditions. The lower number may reflect the frequency of the small THE1a subgroup, of which only three copies not isolated with an o-repeat clone are found in the databanks. The higher number, which has generally been adopted as the number of THE1s in the human genome, may include most or all of the closely related MstII elements. Frequency of the latter group has been estimated to be only 4–8,000 (18, 21) using probes that lack the (best conserved) terminal bases.

The only frequency information available for the rodent elements comes from the observation that hybridization of nick-translated total mouse genomic DNA was as strong to a 200 bp MT-fragment as to clones carrying the 130–150 bp B1 and B2 SINEs (26). B1 and B2 each have an estimated frequency of 80,000 elements in the mouse genome (1). Correcting for the difference in length, this result predicted about 55,000 MT elements in the mouse genome.

Table 3 lists the recurrence of each family in sequences present in GenBank 71 and an estimate of their frequency in the genomes. The estimates are based on the recurrence of the families relative to each other and to the roughly 1,500 Alu (Jerzy Jurka, pers. commun.), 270 B1 and 160 B2 elements present in the human and mouse entries in this release. The numbers obtained in this way may form an underestimation, since MaLRs are—probably unlike SINEs—underrepresented in introns (Table 2), which form a major part of the available non-coding sequence information. The higher limit of reiteration frequencies in human shown in Table 3 is based on the assumption that 37,000 is the total number of THE1- and MstII-LTRs in the human genome and that maximally 10% of the elements are complete retrotransposons with two LTRs. An even higher estimate (>200,000) would follow from the presence of 25 MaLR-LTR sequences that can be detected in 271 randomly obtained chromosome 4 sequence tags with an average length of 440 bp (data not shown, 56). The random nature with which these sequences were acquired could make this last method of estimation actually the most accurate (especially when more sequence tags become available), unless



**Figure 5.** A putatively orthologous MaLR sequence in mouse and human. **a**) Comparison of the mouse and human immunoglobulin heavy chain  $C_\mu$ - $C_\delta$  intergenic region aligned relative to the putatively orthologous MLT1. In the mouse this MaLR has largely been deleted and the remainder has been inverse duplicated. MaLR-LTRs are indicated with open arrows,  $C_\mu$  and  $C_\delta$  exons are indicated with boxes,  $\mu_m$  = membrane carboxyl-terminal exons for  $C_\mu$ . **b**) Alignment of the human (HUMIGMUD, 49) and mouse (MUSIGMUD3, 51) orthologous sequences and the MLT1c-LTR consensus sequence. Bars indicate identical nucleotides between the MLT1c-consensus or the human element and either one of the mouse copies. Flanking the LTR, the 5' end of the consensus MST internal sequence (a consensus MLT1 internal sequence can not yet be derived) is shown in lower-case. Evidence for a common origin of the human and mouse elements is three-fold. (i) They have the same location. (ii) The mouse element is more similar to the human 5' LTR than to any other (MLT1c-) sequence in the databases or even the MLT1c-consensus. Several bases shared between human and mouse are different from the consensus sequence, possibly reflecting mutations that took place after integration but before the rodent-primate split. (iii) Sequence similarity extends into the internal sequence, while 90–95% of the MaLRs occur as solitary LTRs with the internal sequence cleanly deleted.

chromosome 4 has an unrepresentative number of MaLRs. The higher limit of the rodent elements' reiteration frequencies is based on the aforementioned hybridization experiments.

Heinlein *et al.* (26) published hybridization results indicating that MTs are much more commonly cotranscribed in mouse brain than B1, B2 or LINEs. However, MaLRs seem not significantly overrepresented in human brain transcripts compared to other repetitive elements; among the 2723 published 'expressed sequence tags' (57) from human brain cDNA libraries are 313 with Alu, 58 with L1, and 48 with MaLR sequences (HUMXT0... in Table 1). These numbers are comparable to the estimated relative numbers of these repeats in the genome. Data on the MaLR sequences present in these expressed sequence tags will be added to the EST database (57).

It should be noted that the newly discovered MLT1-family has almost twice as many representatives in the human genome as the combined THE1/MstII-family. It could very well be that more distantly related MaLR families exist in both the human and mouse genome, and that the total number of MaLRs is significantly higher than is calculated here. The present estimate of 40,000 to 100,000 MaLRs implies that they occur on average each 30 to 100 kb and comprise 0.5% to 2% of both the human and mouse genome. Furthermore, the presence of transcription termination sites and probably other transcriptional regulatory elements in the MaLR-LTRs would suggest that the distribution of MaLRs has had a considerable influence on the evolution of the mammalian genome.

## ACKNOWLEDGMENTS

I thank Arthur D. Riggs, Gerald P. Holmquist and Jerzy Jurka for helpful comments in preparation of this manuscript. This work was supported by an NIH grant (AG08196) and a DOE grant (DE-FG03-91ER61137) to A. D. Riggs. The consensus sequences will be deposited at and available through the National Center for Biotechnology Information.

## REFERENCES

- Deininger, P.L. (1989) In Berg, D.E. and Howe, M.M. (eds), *Mobile DNA*. American Society for Microbiology, Washington, D.C., pp. 619–636.
- Hwu, H.R., Roberts, J.W., Davidson, E.H. and Britten, R.J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 3875–3879.
- Okada, N. (1991) *Curr. Opin. Genet. Devel.* **1**, 498–504.
- Hutchison, C.A.H., Hardies, S.C., Loeb, D.D., Shehee, W.R., and Edgell, M.H. (1989) In Berg, D.E. and Howe, M.M. (eds.), *Mobile DNA*. American Society for Microbiology, Washington, D.C., pp. 593–617.
- Martin, S.L. (1991) *Curr. Opin. Genet. Devel.* **1**, 505–508.
- Leib-Mösch, C., Brack-Werner, R., Werner, T., Bachmann, M., Faff, O., Erfle, V., and Hehlmann, R. (1990) *Canc. Res.* **50**, 5636s–5642s.
- Keshet, E., Schiff, R., and Itin, A. (1991) *Adv. Cancer Res.* **56**, 215–251.
- Doolittle, R.F., Feng, D.F., Johnson, M.S., and McClure, M.A. (1989) *Quart. Rev. Biol.* **64**, 1–30.
- Temin, H.M. (1989) *Nature* **339**, 254–255.
- Xiong, Y., and Eickbush, T.H. (1990) *EMBO J.* **9**, 3353–3362.
- McClure, M.A. (1991) *Mol. Biol. Evol.* **8**, 835–856.
- Brown, P.O. (1990) *Curr. Top. Microbiol. Immunol.* **157**, 19–48.
- Paulson, K.E., Deka, N., Schmid, C.W., Misra, R., Schindler, C.W., Rush, M.G., Kadyk, L., and Leinwand, L. (1985) *Nature* **316**, 359–361.
- Paulson, K.E., Matera, A.G., Deka, N., and Schmid, C.W. (1987) *Nucleic Acids Res.* **15**, 5199–5215.
- Deka, N., Willard, C.R., Wong, E. and Schmid, C.W. (1988) *Nucleic Acids Res.* **16**, 1143–1151.
- Sun, L., Paulson, K.E., Schmid, C.W., Kadik, L., and Leinwand, L. (1984) *Nucleic Acids Res.* **12**, 2669–2690.
- Jurka, J. (1990) *Nucleic Acids Res.* **18**, 137–141.
- Kaplan, D.J., Jurka, J., Solus, J.F., and Duncan, C.H. (1991) *Nucleic Acids Res.* **19**, 4731–4738.
- Jaiswal, A.K., Gonzalez, F.J., and Nebert, D.W. (1985) *Nucleic Acids Res.* **13**, 4503–4520.
- Lawrance, S.K., Das, H.K., Pan, J., and Weissman, S.M. (1985) *Nucleic Acids Res.* **13**, 7515–7528.
- Mermer, B., Colb, M., and Krontiris, T.G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3320–3324.
- Fields, C.A., Grady, D.L., and Moyzis, R.K. (1992) *Genomics* **13**, 431–436.
- Law, M.L., Xu, Y.X., Berger, R. and Tuñg, L. (1987) *Som. Cell Molec. Genet.* **13**, 381–389.
- Law, M.L., Gao, J., and Puck, T.T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8472–8476.
- Fisher, E.M., Alitalo, T., Luoh, S.-W. and De la Chapelle, A. (1990) *Genomics* **7**, 625–628.
- Heinlein, U.A.O., Lange-Sablitzky, R., Schaal, H., and Wille, W. (1986) *Nucleic Acids Res.* **14**, 6403–6416.
- Bastien, L., and Bourgaux, P. (1987) *Gene* **57**, 81–88.
- Caddle, M.S., Lussier, R.H., and Heintz, N.H. (1990) *J. Mol. Biol.* **211**, 19–33.
- Wilbur, W.J., and Lipman, D.J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
- Sobel, E., and Martinez, H.M. (1985) *Nucleic Acids Res.* **14**, 363–374.
- Jukes, T.H., and Cantor, C.R. (1969) In: Munro H.N. (ed.) *Mammalian protein metabolism*. Academic Press, New York, pp. 21–123.
- Willard, C. (1987) Ph.D. dissertation, University of California, Davis.
- Fan, W., Trifiletti, R., Cooper, T., and Norris, J.S. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 6104–6108.
- Gonzalez, F.J., and Kasper, C.B. (1983) *J. Biol. Chem.* **258**, 1363–1368.
- Kimura, S., Hanioka, N., Matsunaga, E. and Gonzalez, F.J. (1989) *DNA* **8**, 503–516.
- Hollis, G.F., Hietr, P.A., McBride, O.W., Swan, D., and Leder, P. (1982) *Nature* **296**, 321–325.
- Birnsteil, M.L., Busslinger, M., and Stub, K. (1985) *Cell* **41**, 349–359.
- Levin, H.L., and Boeke, J.D. (1992) *EMBO J.* **11**, 1145–1153.
- Wang, G.H., and Seeger, C. (1992) *Cell* **71**, 663–670.
- Deininger, P.L., Batzer, M.A., Hutchinson, C.A., and Edgell, M.H. (1992) *Trends Genet.* **8**, 307–311.
- Sekiguchi, T., Nohiro, Y., Nakamura, Y., and Nishimoto, T. (1991) *Mol. Cell. Biol.* **11**, 3317–3325.
- Maric, S.C., Crozat, A., and Jänne, O.A. (1992). *J. Biol. Chem.* **267**, 18915–18923.
- Bulmer, M., Wolfe, K.H., and Sharp, P.M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 5974–5978.
- O'hUigin, C., and Li, W.H. (1992) *J. Mol. Evol.* **35**, 377–384.
- Schmid, C.W., Wong, E.F.K., and Deka, N. (1990) *J. Mol. Evol.* **31**, 92–100.
- Bailey, W.J., Fitch, D.H.A., Tagle, D.A., Slightom, J.L., and Goodman, M. (1991). *Mol. Biol. Evol.* **8**, 155–184.
- Tagle, D.A., Stanhope, M.J., Siemieniak, D.R., Benson, P., Slightom, J.L., and Goodman, M. (1992) *Genomics* **13**, 741–760.
- Li, W.H., Tanimura, M., and Sharp, P.M. (1987) *J. Mol. Evol.* **25**, 330–342.
- Milstein, C.P., Deverson, E.V., and Rabbits, T.H. (1984) *Nucleic Acids Res.* **12**, 6523–6535.
- Word, C.J., Blattner, F.R., and Kuziel, W.A. (1989) *Int. Immunol.* **1**, 296–309.
- Richards, J.E., Gilliam, A.C., Shen, A., Tucker, P.W., and Blattner, F.R. (1983) *Nature* **306**, 483–487.
- Akahori, Y., Handa, H., Imai, K., Abe, M., Kameyama, K., Hibiya, M., Yasui, H., Okamura, K., Naito, M., Matsuo, H., and Kurosawa, Y. (1988) *Nucleic Acids Res.* **16**, 9497–9511.
- Lloyd, J.A., Lamb, A.N., and Potter, S. (1987) *Mol. Biol. Evol.* **4**, 85–98.
- Ryan, S.C., and Dugaiczak, A. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9360–9364.
- Pascale, E., Valle, E., and Furano, A.V. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9481–9485.
- Goold, R.D., diSibio, G., Dugaiczak, A., Smith, K.A., Xu, H., Magrane, G., Lang, D.B., Cox, D.R., Masters, S.B., and Myers, R.M. (unpublished) HUM4STS...
- Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., and Venter, J.C. (1992) *Nature* **355**, 632–634.
- Li, W.H., Gouy, M., Sharp, P.M., O'hUigin, C., and Yang, Y.W. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 6703–6707.
- Easteal, S. (1992) *Bioessays* **14**, 415–419.
- Bodrug, S.E., Ray, P.N., Gonzalez, I.L., Schmickel, R.D., Sylvester, J.E., and Worton, R.G. (1987) *Science* **237**, 1620–1623.